Cross-linguistic network structure effects on nonword acceptability judgements

Philippa Shoemark

3861047

Supervisor: Dr. James Kirby

Date of submission: April 3, 2013

Word count: 7110

University of Edinburgh
School of Philosophy, Psychology and Language Sciences
Contents

1 Introduction 2

1.1 Nonword acceptability judgements ....................... 2
1.2 Cross-linguistic variation ............................... 6
1.3 Network Science ....................................... 9

2 Method 11

2.1 Materials ............................................. 11
2.1.1 English Materials .................................. 11
2.1.2 Cantonese Materials ............................... 12
2.2 Networks ............................................ 12

3 Results 13

3.1 Multiple linear regression analyses ....................... 14
3.1.1 Cantonese network with tone instantiated as a segment . 14
3.1.2 Cantonese network with tone ignored ................. 15
3.1.3 English network constructed from the complete lexicon . 16
3.1.4 English network constructed from monosyllable lemmas . 17
3.2 Assumptions of multiple linear regression analyses ........ 18
3.2.1 Normally distributed residuals .................... 18
3.2.2 Homoscedasticity ................................ 19
3.2.3 Linearity ........................................ 23
3.2.4 Lack of multicollinearity .......................... 24
3.3 Summary ............................................ 30

4 Discussion 31

5 Conclusion 36
1 Introduction

1.1 Nonword acceptability judgements

Speakers of a given language are intuitively able to distinguish between sound sequences which happen not to be words in their language but conceivably could be, and others which could not.

In early generative phonology, a tripartite categorical distinction was drawn between occurring word forms, such as /brIk/ in English; non-occurring but admissible forms, such as /blIk/, which became known as accidental gaps; and non-occurring but inadmissible forms, such as /bnIk/, which became known as systematic gaps. In order to account for how a child who hears /brIk/ but not /blIk/ or /bnIk/ learns to categorise /blIk/ as an accidental gap and /bnIk/ as a systematic gap, Chomsky and Halle [8] posited that the child constructs a grammar containing an intricate set of ordered rules which generate /blIk/ but are violated by /bnIk/.

However, studies by Greenberg and Jenkins [15] and Scholes [24], in which participants were asked to rate the acceptability of nonwords on a gradient scale, revealed that speakers are able to make scalar judgements about well-formedness as opposed to the strictly categorical ones predicted by the early generative phonology model. Scholes [24] and Chomsky and Halle [9] attempted to incorporate these gradient distinctions into models using feature-based grammatical rules, while Greenberg and Jenkins took a different approach, arguing that scalar nonword acceptability judgements can be accounted for by looking at the similarity of the whole word form to existing word forms in the lexicon, without the need for grammatical rules.

Subsequent studies have consistently found nonword acceptability judge-
ments to be gradient in nature - in fact Hayes and Wilson [17, p. 382] state that gradient judgements “have been found in every experiment that allowed participants to rate forms on a scale”. Despite concerns that fine-grained distinctions in non-word acceptability might be more representative of experimental noise than of participants’ genuine preferences, Albright [1, p. 28] argues that gradient judgements are robust, as participants in his experiment “show a relatively high degree of between-subject agreement, and the differences in mean ratings are replicable”.

There have been numerous attempts to model the mechanisms behind gradient nonword acceptability judgements, but no definitive answer has been found as to whether the gradience in judgements is better explained by phonotactic knowledge - speakers’ ability to decompose new words into segments and assess the probability or legality of the sequence of segments; or by lexical similarity - the degree of overlap between the word-form as a whole and other word-forms stored in the lexicon.

One simple metric for lexical similarity that has consistently been shown to correlate with nonword acceptability judgements is neighbourhood density - a measure of the number of phonological neighbours a given word form has. If we model the lexicon of phonological forms as a network in which nodes correspond to word forms, and an edge connects two nodes if and only if the Levenshtein edit distance between them is 1 (that is to say, if one can be transformed into the other via the insertion, deletion or substitution of a single phoneme), then the neighbourhood density of a particular word form is equal to its degree - i.e. the number of edges connected to it. For example, there are nine English words that you can get to from the word fish by inserting, deleting or substituting a single phoneme, while there are only three words you can get to from the word
squid, as illustrated in Figure 1

![Diagram showing neighbourhood density for fish and squid](image)

Figure 1: The word *fish* has a neighbourhood density of 9 in English, while the word *squid* has a neighbourhood density of 3.

With simple neighbourhood density, there is a sharp distinction between neighbours and non-neighbours, so that all words which differ from a given word by more than one phoneme are counted as equally dissimilar to that word, and all words within a single phoneme edit distance are counted as equally similar. For example, since neither the word *frog* nor the word *chimp* can be changed into the word *dog* via a single phoneme addition, deletion or substitution, the word *frog* is considered by a simple neighbourhood density model to be just as dissimilar to the word *dog* as the word *chimp* is. More nuanced metrics of lexical similarity have been proposed, such as Bailey and Hahn’s [5] Generalized Neighborhood Model, in which all word-forms are considered to be neighbours of all other word-forms, just to varying degrees depending upon the distance between them. The Generalized Neighborhood Model uses a weighted edit distance dependent upon the number of features shared by two phonemes, and favours substitutions over insertions and deletions, as well as taking the token frequency of neighbouring words into account.

A simple model of phonotactic probability that has consistently been shown
to correlate with non-word acceptability is transitional bi-gram probability. The probability of a pair \( ab \) of adjacent phonemes occurring together (or more precisely, the probability of phoneme \( b \) occurring immediately after phoneme \( a \)) is calculated by dividing the number of times the pair \( ab \) occurs in a corpus by the number of times \( a \) occurs in total. The phonotactic probability of the entire word-form is the aggregate probability of its component bi-phones. Models of phonotactic probability incorporating features [1] and prosodic structure [13, 26, 5] have also been proposed.

While lexical similarity and phonotactic probability both correlate highly with English speakers’ nonword acceptability judgements, they are also correlated with each other, making it difficult to be sure that one mechanism is not implicit in the other. Bailey and Hahn [5] set out to tease apart the individual contributions of the two mechanisms by directly comparing them on the same task, and found that while lexical similarity was the more important predictor, it did not completely subsume the effect of phonotactic probability, suggesting that speakers use a combination of both mechanisms.

Albright made a similar comparison using different data, and found a primarily phonotactic effect - the opposite of Bailey and Hahn’s result. One possible explanation for this discrepancy is Shademan’s [25] hypothesis that the balance between the two mechanisms is swayed in favour of lexical similarity by the inclusion of real-world stimuli. This is in keeping with Vitevitch and Luce’s [28] findings that neighbourhood density affects word recognition “when the lexical level is invoked using real words” [28, p. 325], but that in the absence of real world stimuli, effects of phonotactic probability emerge.

After analysing the unique contributions of phonotactic probability and lexical similarity, Bailey and Hahn [5, p. 570] noted that there was still a con-
siderable amount of variance in participants’ acceptability judgements left unexplained, which, they argued, “points to limitations in present conceptions of either phonotactics or lexical influence”. In order to address this, they extensively explored potential improvements that could be made to both types of model, and concluded that “although modest improvements can be made in phonotactic measures, [...] more important contributions will likely come from better models of lexical neighbourhoods”, since with lexical neighbourhoods “there is much more to explore and [...] much less ground has been covered by previous research”.

1.2 Cross-linguistic variation

All of the nonword acceptability studies discussed so far have focused on judgements made by English speakers. Few studies have investigated the influence of phonotactic probability and lexical similarity on acceptability judgements in other languages, but there have been studies on Mandarin and Cantonese, two languages whose phonotactics are very different from those of English.

Kirby and Yu [18] found that neighbourhood density was more important as a predictor of nonword acceptability for Cantonese than other studies had reported it to be for English, while the effect of phonotactic probability was relatively weak in Cantonese. They suggested that this might be “an effect of the relative phonological densities of the Cantonese and English lexica” [18, p. 1389].

The potential phonotactic space of Cantonese is much smaller than that of English, since Cantonese syllable structure is more restricted. In addition to its tone, a syllable in Cantonese consists of between one and three phonemes: an optional onset, an obligatory nucleus, and an optional coda. Thus, there are five
ways in which Cantonese syllables can be structured in terms of vowels and con-
onsonants: CVC, CV, VC, V, and C (syllabic consonant). The onset and coda are
also optional in English syllables, and since English allows complex onsets and
codas consisting of up to three or four consonants each (e.g. /splæt/, /krɪspr/,
/strɛŋkts/), the number of possible configurations of vowels and consonants in
English syllables is much greater. According to Kirby and Yu, nearly a third of
the space of possible Cantonese syllables\(^1\) is occupied by lexical items, whereas
in English, lexical items occupy a far smaller proportion of the potential phono-
tactic space. Hence in Cantonese, most nonwords have at least one neighbour
within an edit distance of 1, but in English a nonword like /drɛsp/ can have no
neighbours yet still be judged to be highly acceptable by English speakers.

It is worth noting that the apparent discrepancy between the relative impor-
tance of neighbourhood density in English and in Cantonese could also be due
at least in part to the fact that Kirby and Yu’s experiment included real word
stimuli as well as nonwords, whereas no real words were included in Frisch et al’s
study [13], which Kirby and Yu cite as an example of a study in which phono-
tactic probability emerges as a relatively strong predictor of English nonword
acceptability. As mentioned above, there is evidence to suggest that neighbour-
hood density has a stronger influence on the processing of nonwords when real
word stimuli are also presented than when they are not.

Nevertheless, a closer look at Frisch et al’s results lends support to Kirby

\(^1\)Syllables generally correspond to words in Cantonese, although some syllables do not
carry meaning individually, and some words are formed by compounding two or more syll-
ables. Alpatov [3, p. 1803] argues that syllables in Chinese and other East and South East
Asian languages are “the most important psycholinguistic units kept to the memory of na-
tive speakers” and “[in this respect they can be compared with the words of the European
languages”.
and Yu’s explanation. Although Frisch et al found that the effect of phonotactic probability was stronger than that of neighbourhood density overall, there was no significant difference between phonotactic probability measures and neighbourhood density for nonwords with high probability constituents. The weaker overall effect of neighbourhood density was due to the fact that for nonwords with low probability constituents, neighbourhood density did not correlate very strongly with participants’ judgements because many of these nonwords had no neighbours at all, resulting in a floor effect. In Cantonese, since most nonwords have at least one neighbour in the lexicon, there is no floor effect for stimuli with low probability constituents, and so it makes sense that neighbourhood density fares comparatively better overall.

Myers and Tsay [21] found that in Mandarin, which also has a small, dense phonological lexicon, neighbourhood density was positively related to judgements about the typicality of real Mandarin words, but negatively related to judgements of nonwords. They speculated that while English speakers judge nonwords with a high level of similarity to other words in the lexicon as particularly acceptable, Mandarin speakers might instead interpret them as “a misleading illusion that must be suppressed”[21, p. 33].

While there definitely seem to be cross-linguistic difference in the extent to which and the ways in which lexical measures influence nonword acceptability judgements, and it seems plausible that these differences may have something to do with cross-linguistic differences in the structure of the phonological lexicon, a better understanding of these structures is needed in order to fully explain the observed effects.
1.3 Network Science

Network science is concerned with analysing the structure of complex systems in order to gain insights into how such systems function and evolve, and has been applied to a wide variety of physical, biological and social phenomena.

Recent studies ([27], [4], [6], [7]) have begun to explore what network science can tell us about the structure of the phonological lexicon, and how that structure influences the processing of phonological forms. Arbesman et al [4] compared the phonological network structures of five languages from different language families. Certain distinctive structural properties (such as small-world characteristics and assortative mixing by degree) are exhibited by all five languages, and Arbesman et al posit that these properties may facilitate rapid and efficient search through the lexicon. But there are also considerable differences between the network structures of different languages - for example, the average clustering coefficient in the largest component in the Mandarin network is twice that of the Spanish network.

A node’s clustering coefficient is a reflection of how many of its neighbours are also neighbours of each other. Clustering coefficient is defined as the number of edges between a node’s neighbours divided by the total number of possible edges between them. Values of clustering coefficient range from 1, if a node’s neighbourhood is completely interconnected, to 0, if none of a node’s neighbours are connected to each other. For example, Figure 2 shows three words with the same neighbourhood density but different clustering coefficients. Strive, whose neighbours are all interconnected, has a clustering coefficient of 1. Drink, whose neighbours have three edges between them out of a possible six, has a clustering coefficient of 0.5. And trend, whose neighbours have no connections between them, has a clustering coefficient of 0.
Figure 2: Three words with the same neighbourhood density but different clustering coefficients.
Chan and Vitevitch found that clustering coefficient affects the speed and accuracy of both spoken word recognition [6] and spoken word production [7]. Compared to words with a high clustering coefficient, words with a low clustering coefficient were more accurately identified in a perceptual identification task, more quickly responded to in both a lexical decision task and in a picture naming task, and occurred less frequently in a corpus of speech errors. These findings suggest that lexical access is influenced not only by the number of neighbours a target word has, but also by the similarity relationships among its neighbours. Clustering coefficient therefore seems like a promising avenue to explore in the search for more nuanced models of lexical neighbourhoods that might account for as-yet unexplained variance in nonword acceptability judgements.

2 Method

2.1 Materials

The present dissertation makes use of stimuli, acceptability judgements, and segmental bigram probabilities from previous studies on non-word acceptability.

2.1.1 English Materials

The English stimuli and acceptability judgements were collected by Albright and Hayes [2] in order to obtain baseline phonological well-formedness scores for their study on how speakers inflect novel word forms. Albright and Hayes presented 90 monosyllabic nonwords to adult native speakers of American English. The stimuli were presented aurally, first in isolation and then in a carrier sentence, and participants were asked to rate each non-word on a 7-point Likert scale based on how natural or English-like it sounded.
The segmental bigram probabilities for 87 of Albright and Hayes's stimuli were calculated by Gorman [14]. Three of the original stimuli were real English words and have been excluded from Gorman’s analysis, and thus from the present analysis also.

2.1.2 Cantonese Materials

The Cantonese stimuli, acceptability judgements and segmental bigram probabilities are from a follow-up study [19] to the above-mentioned paper by Kirby and Yu [18].

73 real Cantonese words and 143 nonwords, all monosyllabic with CV(C) structure, were presented aurally to native speakers of Cantonese, who were asked to rate their wordlikeness on a 7-point Likert scale. The real words were not included in the present analysis.

2.2 Networks

Phonological networks, in which nodes represent phonological word-forms, and an edge connects any two nodes with a Leviensthein edit distance of 1, were constructed from English and Cantonese lexicons.

Two networks were constructed for Cantonese: one in which tone was effectively instantiated as another segment appended to the end of each word; and one in which tone was ignored. The Cantonese lexicon contains 1884 unique syllables, and is derived from the Chinese Character Database by the Chinese University of Hong Kong [11].

A network for English was constructed from the CMU Pronouncing Dictionary [10], which contains 114,966 unique word-forms. A separate English network containing only monosyllabic words was also constructed, since the
nonword stimuli are monosyllabic and it’s possible that participants might consider the degree of support given just by English monosyllables, rather than the entire lexicon of English word-forms. Furthermore, as the Cantonese network contains only monosyllables, restricting the English network to monosyllables may make for a fairer comparison. The lexicon of English monosyllables, which contained 3653 unique word-forms, was obtained by cutting down the set of CMU Pronouncing Dictionary entries to only those entries that also occur in the set of monosyllable lemmas in CELEX [22].

Nonwords were inserted one at a time into the phonological networks for the relevant language, and the NetworkX Python Package [16] was used to calculate the clustering coefficients and neighbourhood densities of each nonword.

3 Results

Clustering coefficient is meaningless for nodes which have one or zero neighbours, since by definition it computes the likelihood that any two neighbours of a given node will also be neighbours of each other (recall that it is calculated by dividing the number of edges between a node’s neighbours by the number of possible edges between them - so if there are zero or one neighbours, the clustering coefficient is zero divided by zero, which is undefined). Nonwords which have a neighbourhood density of less than two have therefore been excluded from the analyses. No nonwords had neighbourhood density less than two using the Cantonese network in which tone was ignored, but 4 such items were excluded from the results of the Cantonese network in which tone was instantiated as a segment, 13 from the results of the network built from the full English lexicon and 22 from the results of the English monosyllable network.
3.1 Multiple linear regression analyses

Multiple linear regression analyses were carried out on the data from each network, in order to evaluate the independent contributions of bigram probability, neighbourhood density and clustering coefficient in accounting for the variance in nonword acceptability ratings. In order to make their distributions closer to the normal probability distribution so that they would better fit the assumptions underlying regression, the clustering coefficient variables were log transformed and the neighbourhood density variables were square-root transformed prior to the analyses.

3.1.1 Cantonese network with tone instantiated as a segment

A partial model using neighbourhood density and bigram probability as predictors accounted for around a quarter of the variance in non-word acceptability judgements ($R_{adjusted}^2 = 0.2527$, $F(2,131) = 23.49$, $p < 0.001$). As can be seen in Table 1, neighbourhood density was significant controlling for bigram probability, as was bigram probability controlling for neighbourhood density.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$-value</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.302</td>
<td>4.026</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.430</td>
<td>5.725</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 1: Cantonese with tone - partial model

Including clustering coefficient in the model provided a small but significant improvement ($F_{for-change}(1,130) = 7.853$, $p < 0.01$). As Table 2 shows, clustering coefficient has the smallest effect of the three predictors, but is nevertheless significantly related to nonword acceptability ratings when the other predictors are held constant. The sign of the $\beta$ value is positive, meaning non-
words with higher clustering coefficients are rated more highly than nonwords with lower clustering coefficients. The overall model fit was $R^2_{\text{adjusted}} = 0.2899$, $F(2,131) = 19.1$, $p < 0.001$.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$-value</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.443</td>
<td>4.991</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.391</td>
<td>5.251</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Log Clustering Coefficient</td>
<td>0.253</td>
<td>2.802</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Table 2: Cantonese with tone - full model

### 3.1.2 Cantonese network with tone ignored

All three predictors performed better when tone was ignored. The partial model using just neighbourhood density and bigram probability accounted for considerably more variance ($R^2_{\text{adjusted}} = 0.3909$, $F(2, 140) = 46.57$, $p < 0.001$) than when tone was instantiated as a segment, with each predictor emerging as significant when controlling for the other, as shown in Table 3.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$-value</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.482</td>
<td>7.273</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.507</td>
<td>7.612</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 3: Cantonese without tone - partial model

Table 4 shows that clustering coefficient was also significant when added to the model, and again had a positive $\beta$-value. This time, the magnitude of its individual effect was similar to that of bigram probability, and its inclusion accounted for a further 16% of the variance in the ratings ($R^2_{\text{adjusted}} = 0.5543$, $F(2, 140) = 50.47$, $p < 0.001$).
\( F(2, 140) = 55.83, p < 0.001 \),

<table>
<thead>
<tr>
<th>Predictor</th>
<th>( \beta )-value</th>
<th>( t )-value</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.637</td>
<td>10.464</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.419</td>
<td>7.197</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Log Clustering Coefficient</td>
<td>0.448</td>
<td>7.234</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 4: Cantonese without tone - full model

3.1.3 English network constructed from the complete lexicon

Neighbourhood density and bigram probability together accounted for around 45% of the variance in English nonword acceptability ratings \( R_{\text{adjusted}}^2 = 0.4464 \), \( F(2,131) = 30.03, p < 0.001 \), and the inclusion of clustering coefficient made a small but significant improvement \( R_{\text{adjusted}}^2 = 0.4954 \), \( F\text{-for-change}(1,69) = 7.7918, p < 0.01 \). In contrast to the Cantonese results, the relationship between clustering coefficient and acceptability rating was negative.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>( \beta )-value</th>
<th>( t )-value</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.326</td>
<td>2.364</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.395</td>
<td>2.860</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Table 5: Complete English lexicon - partial model

Comparing Table 5 and Table 6, we see that while bigram probability was significant when controlling for neighbourhood density alone, it was no longer significant when clustering coefficient was added to the model. This indicates that the effect of bigram probability was completely subsumed by the combined effects of neighbourhood density and clustering coefficient.
3.1.4 English network constructed from monosyllable lemmas

When the values of neighbourhood density and clustering coefficient were obtained from the network consisting of only monosyllabic English lemmas, all three predictors had slightly weaker effects. Although the same bigram probability values were used in both English datasets, more nonwords had neighbourhood densities of less than 2 in the monosyllable network than in the network built from the complete English lexicon, and so more items were excluded from the analysis, which resulted in a weaker relationship between bigram probability and acceptability rating.

The partial model using just neighbourhood density and bigram probability accounted for 22% of the variance in non-word acceptability ratings. As Table 7 shows, bigram probability was not significant controlling for neighbourhood density.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>β-value</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.330</td>
<td>2.030</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.220</td>
<td>1.232</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 7: English monosyllable lemmas - partial model

The inclusion of clustering coefficient did not significantly improve the model
\( R^{2}_{\text{adjusted}} = 0.2425, F\text{-for-change}(1,60) = 2.7106, p \text{ n.s.} \), however the regression coefficients for the full model are still shown in Table 8, as although the effect of clustering coefficient did not reach significance, it is still worth noting that the sign of its \( \beta \)-value is negative. We can be sure that if there is a relationship between clustering coefficient and nonword acceptability ratings in English, it is negative, whereas in Cantonese it is positive.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>( \beta )-value</th>
<th>( t )-value</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Root Neighbourhood Density</td>
<td>0.420</td>
<td>2.479</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Bigram Probability</td>
<td>0.065</td>
<td>0.363</td>
<td>n.s.</td>
</tr>
<tr>
<td>Log (Clustering Coefficient + 1)</td>
<td>-0.203</td>
<td>-1.646</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 8: English monosyllable lemmas - full model

3.2 Assumptions of multiple linear regression analyses

The multiple linear regression models described above tell us how much of the variance in the nonword acceptability ratings collected by Albright & Hayes and Kirby et al is accounted for by clustering coefficient, neighbourhood density and bigram probability. In order to be able to generalise these models from the observed judgements to judgements for other nonwords, we need to ascertain whether the underlying assumptions of multiple linear regression are valid.

3.2.1 Normally distributed residuals

One assumption underlying multiple linear regression is that the residuals are approximately normally distributed. Histograms showing the distribution of studentized residuals for each of the four models are displayed in Figure 3. We can see from these histograms that the residuals in both of the Cantonese
models are positively skewed, and the residuals in the model for the English network constructed from the complete lexicon are negatively skewed. However, according to Fox [12, p. 40], “the levels of tests and confidence intervals are approximately correct in large samples even when the assumption of normality is violated”, so the fact that the residuals in three of the models are non-normally distributed is not too much of a cause for concern. Fox does note that non-normality can markedly decrease the efficiency of least-squares estimation for error distributions with heavy tails, but the normal-quantile comparison plots in Figure 4 indicate that the distributions are not very heavily tailed, as in the English models the tails stay within the 95 percent confidence envelope around the fitted line, and in the Cantonese model they only just stray outside.

### 3.2.2 Homoscedasticity

Another assumption of multiple linear regression is homoscedasticity - the assumption that the variance in the residuals is constant across all levels of the predictors. Plots of residuals against fitted values for each model are shown in Figure 5. In all four of the plots, the variance in the residuals depends on the fitted value, with greater variability in the residuals when the estimated ratings are higher. Thus, the assumption of homoscedasticity has been violated in all models. According to Fox [12, p. 49], “the least-squares estimator is unbiased and consistent even when the error variance is not constant”, but “its efficiency is impaired and the usual formulas for standard error are inaccurate.”. So while the coefficients of the regression models are robust against violations of the homoscedasticity assumption, the significance tests are not, meaning we can trust that the models are a good fit to the observed ratings, but not necessarily that they are generalisable to ratings for other non-words.
Figure 3: Distribution of studentized residuals for the four models
Figure 4: Normal quantile-comparison plots for the four models
Figure 5: Residuals vs Fitted Values for the four models
3.2.3 Linearity

If the relationship between a predictor variable and the outcome variable is not linear, a multiple linear regression model will not capture the relationship very accurately. Figures 6-9 display partial-residual plots for each model, which show the relationships between the ratings and each predictor, controlling for the other predictors. Overlaid on each plot is a least-squares line and a Lowess (locally weighted scatterplot smoothing) curve. Comparing the least-squares line and the Lowess curve helps us to assess whether or not a relationship really is linear. For the Cantonese model without tone, it looks like all three predictors are related to ratings in a fairly linear way, but for the Cantonese model with tone, while bigram probability and neighbourhood density are linearly related to ratings, clustering coefficient looks like it would be better approximated by a cubic function. Turning to the English model using the complete lexicon, there is a clear outlier in the bottom left corner of the partial-residual plot for clustering coefficient in figure 8, but aside from that the relationship between clustering coefficient and ratings is reasonably linear. The outlier, which has a much lower rating than expected for its clustering coefficient, is /ploonθ/. It has a clustering coefficient of zero and a neighbourhood density of two, but one of its two neighbours is /ploon/, which, despite having an entry in the CMU pronouncing dictionary, is not, as far as I am aware, a real word of English. If we were to remove /ploon/ from the network on account of its not being a real English word, then /ploonθ/ would only have one neighbour, /plnθ/, and its clustering coefficient would be undefined. Without /ploonθ/, the relationship between clustering coefficient and rating is fairly linear.

In the English model using monosyllable lemmas, the relationship between clustering coefficient and rating is fairly linear for values of clustering coefficient
that are between 0 and 1, but when clustering coefficient is equal to zero or one, the ratings don’t follow the linear trend. However, removing items with clustering coefficients of zero or one does not make a great deal of difference to the slope of the least-squares line, as can be seen by comparing figures 9 and 10.

3.2.4 Lack of multicollinearity

Multicollinearity among the predictor variables is problematic because it increases the variances of the regression coefficients and makes it difficult to interpret the relative importance of individual predictors. The three predictor
Figure 7: Partial-residual plots for Cantonese without tone
Figure 8: Partial-residual plots for English (complete lexicon)
Figure 9: Partial-residual plots for English (monosyllable lemmas)
Figure 10: Partial-residual plots for English (monosyllable lemmas), excluding nonwords with clustering coefficients of 0 or 1.
<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese network with tone</td>
<td></td>
</tr>
<tr>
<td>Square root neighbourhood density</td>
<td>1.48</td>
</tr>
<tr>
<td>Bigram probability</td>
<td>1.04</td>
</tr>
<tr>
<td>Log clustering coefficient</td>
<td>1.52</td>
</tr>
<tr>
<td>Cantonese network without tone</td>
<td></td>
</tr>
<tr>
<td>Square root neighbourhood density</td>
<td>1.18</td>
</tr>
<tr>
<td>Bigram probability</td>
<td>1.08</td>
</tr>
<tr>
<td>Log clustering coefficient</td>
<td>1.22</td>
</tr>
<tr>
<td>English network: complete lexicon</td>
<td></td>
</tr>
<tr>
<td>Square root neighbourhood density</td>
<td>2.48</td>
</tr>
<tr>
<td>Bigram probability</td>
<td>2.98</td>
</tr>
<tr>
<td>Log (clustering coefficient + 1)</td>
<td>1.43</td>
</tr>
<tr>
<td>English network: monosyllables</td>
<td></td>
</tr>
<tr>
<td>Square root neighbourhood density</td>
<td>2.39</td>
</tr>
<tr>
<td>Bigram probability</td>
<td>2.70</td>
</tr>
<tr>
<td>Log (clustering coefficient + 1)</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Table 9: Collinearity diagnostics
variables in each model were tested for collinearity using the Variance Inflation Factor (VIF), which measures the degree to which the variance of the regression coefficient for a particular predictor is inflated by its collinearity with the other predictors in the model. According to Miles and Shevlin [377, p. 130],

“The increase in [the] standard error [of the variable] is equal to the square root of the VIF: when the VIF is equal to four the square root is doubled ($\sqrt{4} = 2$), and so four is often used as an arbitrary cut-off to determine when collinearity has become too serious”

As shown in Table 9, in all four models, all predictors had VIFs well below 4, so we can conclude that collinearity is not a problem for these models.

3.3 Summary

The multiple linear regression analyses revealed that clustering coefficient is positively related to nonword acceptability ratings in Cantonese and negatively related to nonword acceptability ratings in English. A greater amount of the variance in the Cantonese ratings was captured when tone was ignored in the phonological network ($R^2_{\text{adjusted}} = 0.5543$) than when it was instantiated as a segment ($R^2_{\text{adjusted}} = 0.2899$), and a greater amount of the variance in the English ratings was captured when the phonological network was built from the complete lexicon ($R^2_{\text{adjusted}} = 0.4464$) than when only monosyllable lemmas were used ($R^2_{\text{adjusted}} = 0.2425$). The underlying assumptions of multiple linear regression were not met very well, however, so significance values for the individual predictors should not be taken at face value. Since multiple linear regression is fairly robust against violations of the normally distributed residuals and homoscedasticity assumptions, and since the VIF values do not indicate any collinearity problems, the standardised regression coefficients can still give
us some indication of the individual contributions of each predictor, but it is unclear how generalisable the results reported here are. Since the models best fit the data when the clustering coefficient and neighbourhood density values were obtained from the Cantonese network ignoring tone and the English network constructed from the complete lexicon, the discussion will focus on those results.

4 Discussion

The results suggest that clustering coefficient influences nonword acceptability judgements in both languages, but in opposite ways. In Cantonese, non-words with higher clustering coefficients are judged to be more acceptable than non-words with lower clustering coefficients, while in English there is an inverse relationship between clustering coefficients and acceptability judgements.

Chan and Vitevitch [6, 7] explain the effect of clustering coefficient on spoken word recognition and production in English by proposing a framework in which lexical retrieval is viewed as a search through a network, on analogy with search algorithms used to retrieve information from the World Wide Web. Within this framework, an input signal activates a node corresponding to the perceived or intended word-form, and activation spreads from the initially activated node to its neighbours, and then to its neighbours’ neighbours (which includes some activation spreading back to the initially activated node itself), and so on throughout the network, until the activation (a limited cognitive resource) runs out. If the initially activated node has a high clustering coefficient, its neighbours are highly interconnected and so most of the activation bounces back and forth between them rather than dispersing to other parts of the network. If the initially activated node has a low clustering coefficient, the activation spreads further,
as less of it gets used up reverberating between the initial node’s neighbours. The initially activated node will always be highly activated, as it receives direct activation from the input signal in addition to the activation that spreads back to it from its neighbours. High clustering coefficients result in a small local region of the network becoming highly activated, making it difficult for the initially activated node to stand out against its neighbours, whereas low clustering coefficients result in low level activation spreading thinly across the network, making it easy for the initially activated node to stand out. Thus, words with low clustering coefficients can be retrieved more quickly and accurately than words with high clustering coefficients.

If we apply this framework to the problem of assessing non-word acceptability, we could posit that the amount of lexical support a non-word receives is analogous to the percentage of nodes in the network that its activation spreads to, or alternatively we could posit that it is analogous to the amount of activation its immediate neighbours receive. If the assessment of lexical support is based on the percentage of nodes in the whole network that are activated, this explains the inverse relationship between clustering coefficient and non-word acceptability observed in English, since word-forms with low clustering coefficients result in activation being dispersed widely throughout the network, and word-forms with high clustering coefficients result in activation being confined to a small local region. If the assessment of lexical support is based on how highly activated a non-word’s immediate neighbours are, this explains the positive relationship between clustering coefficient and non-word acceptability observed in Cantonese, since word-forms with low clustering coefficients result in activation being spread out thinly whereas word-forms with high clustering coefficients result in a high amount of activation circulating amongst the
This raises the question of why English speakers would assess lexical support differently to how Cantonese speakers do. A possible explanation is that primarily, lexical support is assessed on the basis of the percentage of nodes in the whole network that receive activation, but that in Cantonese there is a ceiling effect because its denser structure means that most nonwords activate a high percentage of the nodes in the network, regardless of their clustering coefficient. Since the majority of nonwords receive high lexical support based on the percentage of the nodes they activate, Cantonese speakers might discriminate between them using a secondary measure of lexical support based on the amount of activation a nonword’s immediate neighbours receive.

This hypothesis of course depends on it actually being the case that a high percentage of the Cantonese network is activated by almost any nonword stimulus. A comparison of some of the global characteristics of the Cantonese and English phonological networks, calculated using NetworkX and displayed in Table 10, suggests that this is not an unreasonable assumption to make. First of all, the Cantonese network has only one component - meaning there are no isolated nodes or separate ‘islands’ of nodes that are connected to each other but not to the rest of the network, whereas English has lots of these ‘islands’ - in fact only about half of the words in the English phonological network make up the largest connected component (a.k.a. the ‘giant component’). This means that in Cantonese, it is possible for activation to spread from any node in the network to any other node, but in English, there are some nodes that simply cannot be reached from other nodes, even if the activation never runs out, because there is no path between them.

The average shortest path length of a network tells us the average number
of edges that have to be traversed in order to get from any given node to any
other given node in the network. To get from chalk to cheese, for example, we
can traverse the edge that joins chalk to check, and then the edge that joins
check to cheese. To get from big to small takes a few more steps: big - bill -
ball - mall - small. The average shortest path length can only be computed for
a fully connected network, so only the average shortest path length of the giant
component was calculated for the English network. The average shortest path
length of the Cantonese network is considerably shorter, in fact it’s less than
half of that of the English giant component. Intuitively, it seems that having a
shorter average path length ought to mean a higher proportion of nodes in the
network can be reached with the same limited amount of activation.

<table>
<thead>
<tr>
<th></th>
<th>Cantonese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>640</td>
<td>114966</td>
</tr>
<tr>
<td>Number of nodes (GC)</td>
<td>'</td>
<td>56689</td>
</tr>
<tr>
<td>Number of components</td>
<td>1</td>
<td>38661</td>
</tr>
<tr>
<td>Avg. shortest path length</td>
<td>2.8</td>
<td>-</td>
</tr>
<tr>
<td>Avg. shortest path length (GC)</td>
<td>'</td>
<td>7.6</td>
</tr>
<tr>
<td>Avg. clustering coefficient</td>
<td>0.48</td>
<td>0.16</td>
</tr>
<tr>
<td>Avg. clustering coefficient (GC)</td>
<td>'</td>
<td>0.26</td>
</tr>
<tr>
<td>Network density</td>
<td>0.03683</td>
<td>0.00005</td>
</tr>
<tr>
<td>Network density (GC)</td>
<td>'</td>
<td>0.00018</td>
</tr>
</tbody>
</table>

Table 10: Characteristics of Cantonese and English phonological networks (GC = Giant Component).

The average clustering coefficient (i.e. the mean of the clustering coefficients
of all the nodes in the network) is much higher for the Cantonese network than
for the English network. Since the Cantonese network is fully connected and
the English network has a lots of islands, a fairer comparison might be between
the average clustering coefficient of the Cantonese network and the English
giant component, but even then the value for Cantonese is almost double the
value for English. Newman [23] modelled the spread of epidemics in networks
with different average clustering coefficients, and found that in highly clustered
networks, epidemics spread quickly and easily to almost every node in the giant
component. Newman explains that this is because in a highly clustered network
there are many redundant paths between nodes, so any individual that can
catch the disease (or any word-form that can be activated) is highly likely to be
reached by one route or another. Since the Cantonese network is fully connected
and has a relatively high average clustering coefficient, we can predict that
activation would indeed spread to a high percentage of the network from almost
any nonword.

Network density is another measure of the connectivity in a network, de-
defined as the ratio of the number of edges in the network to the total number
of possible edges between all pairs of nodes. If this definition sounds familiar,
that’s because clustering coefficient is actually the network density of a particu-
lar node's immediate neighbourhood. Network density tells us what proportion
of the total number of possible edges in the network actually exist, while the av-
erage clustering coefficient tells us what proportion of a given node’s neighbours
will be connected to each other on average. It’s a subtle distinction and both
measures are telling us something about how densely connected the network is,
so both have been reported in Table 10. The network density of the Cantonese
network is almost a thousand times greater than that of the English network,
but since English has lots of islands it’s unsurprising that it has proportionally
fewer edges than the fully connected Cantonese graph. Nevertheless, the Cantonese network density is still 200 times greater than that of the English giant component.

There clearly is a considerable difference between the two languages in terms of the connectivity of their phonological networks, and if Newman is right, then the fact that Cantonese network is so much more densely clustered than English does indeed mean that a much higher proportion of the Cantonese lexicon can be activated by any nonword with at least one real word neighbour. Since all nonwords in the dataset did have at least one real word neighbour, it’s quite plausible that all or almost all of them activated a high or even near-maximal percentage of the lexicon, forcing Cantonese speakers to use a secondary measure of lexical support in order to produce gradient acceptability judgements.

5 Conclusion

The results of the present dissertation indicate that clustering coefficient captures some of the gradience in nonword acceptability judgements that is unaccounted for by neighbourhood density and bigram probability. The effect of clustering coefficient on nonword acceptability in English was found to be negative, whereas a positive effect of clustering coefficient was found on Cantonese nonword acceptability judgements. These opposite effects could potentially be explained using a model of activation spreading through a network, like the one proposed by Chan and Vitevitch [] to explain the effect of clustering coefficient on the speed and accuracy of lexical retrieval. I have hypothesised that English speakers might assess the degree of lexical support for a nonword by the proportion of nodes in the phonological network that its activation spreads to, but that this measure is not adequate for Cantonese speakers to make gradient
distinctions since a high proportion of the phonological lexicon is activated by the majority of possible syllables that can be created within the limits of its phonemic inventory and simple syllable structure. Hence, Cantonese speakers might use an alternative measure of lexical support based on the amount of activation received by a nonword’s immediate (i.e. one phoneme edit-distance) neighbours. It would be interesting to test this hypothesis with a simulation of spreading activation in the English and Cantonese phonological networks.

References


